Amendments to the Specification:

1. Please replace the misspelled word "immediatedly" on page 3, line 12, with the amended word "immediately".

2. Please replace the misspelled word "intial" on page 4, line 19, with the amended word "initial".

3. Please replace the text beginning on page 5, line 15:

> Considering that hundreds of other general-purpose clustering algorithms have been described [KAUFMAN and ROUSSEEUW. Finding Groups in Data: An Introduction to Cluster Analysis, Wiley (1990) and references contained therein]

with the following amended text:

> Considering that hundreds of other general-purpose clustering algorithms have been described [KAUFMAN and ROUSSEEUW. Finding Groups in Data: An Introduction to Cluster Analysis, Wiley (1990) and references contained therein; MANNING et al., Chapter 14, "Clustering", in Foundations of Statistical Natural Language Processing, MIT Press (1999)]

4. Please delete the text beginning on page 6, line 9:

> If genes in a cluster are co-regulated by the same transcription factors, it would consequently be more appropriate to cluster genes on the basis of the similarity of their mRNA synthesis rates (which the transcription factors affect directly), rather than total mRNA levels. However, the clustering methods described above have heretofore been applied only to microarray data corresponding to the total amount of mRNA for each gene, which is the net amount of mRNA resulting from each gene's mRNA synthesis, minus the amount of the gene's mRNA that has been degraded. One objective of the present invention is therefore to provide a method for estimating mRNA synthesis rates from measurements corresponding to total mRNA levels, for each of the genes represented on a microarray, for purposes of clustering.

5. Please replace the text beginning on page 6, line 20:

> Whether clustering is performed using total mRNA levels or estimated mRNA synthesis rates, one needs to compare results made with different clustering algorithms, in order to decide which algrothm is most useful for the data under investigation.

with the following amended text:

> When clustering is to be performed, one therefore needs to compare results made with different clustering algorithms, in order to decide which algorithm is most useful for the data under investigation.

6. Please delete the text beginning on page 8, line 17:

> It is therefore another objective of the present invention to produce automatically generated, quantitative indices (figures-of-merit) of the extent to which genes in a cluster are functionally related to one another, based on information within scientific literature concerning genes present on a microarray. Investigators may use the indices to evaluate the functional relatedness of genes in clusters that were made using a particular clustering algorithm, as well as to compare the performance of different clustering algorithms. In so doing, the investigators use the figure-of-merit indices that are generated by the method to evaluate the quality of the clustering algorithms, based solely on the content of the literature about the genes associated with the clusters.

7. Please replace the text beginning on page 9, line 5:

> In one embodiment of the invention, the figure-of-merit indices are calculated by obtaining text in the scientific literature about genes on a microarray (using an original method that is part of the invention);

with the following amended text:

> In the present invention, clusters of genes are characterized automatically by obtaining text in the scientific literature about genes on a microarray (using an original method that is part of the invention);

8. Please delete the text beginning on page 9, line 12:

genes in the clusters. The figureof-merit indices of the method and Indices are then generated by testing the model's ability to classify additional text about system relate to the percentage of times that the tested classifications are made correctly, as compared with classifications performed on text corresponding to genes placed randomly in clusters.

9. Please replace the text beginning on page 10, line 3:

Unlike the present invention, their method does not make use of information from the clustering of microarray data, and it provides no figure of merit for the quality of microarray clustering.

with the following amended text:

Unlike the present invention, their method does not make use of information from the clustering of microarray data.

10. Please delete the text beginning on page 10, line 10:

Unlike the present invention, their method does not provide a figure of merit for the quality of microarray clustering.

11. Please replace the text beginning on page 10, line 11:

Their method also has the disadvantage

with the following amended text:

Their method has the disadvantage

12. Please delete the text beginning on page 11, line 9:

An option of this clustering module is to estimate mRNA synthesis rates from total mRNA measurements, then perform the clustering using the mRNA synthesis rates.

13. Please replace the misspelled word "identifers" on page 11, line 17, with the word "identifiers".

14. Please delete the text beginning on page 11, line 23, most of which continues on page 12:

Another computer program module produces a quantitative index of the relatedness of genes within each cluster, by testing the model's ability to classify additional text about genes in the clusters. In this embodiment, a figure-of-merit index, which is generated by the method and system, relates to the percentage of times that the test classifications are made correctly, as compared with classifications performed on text that had been clustered randomly.

In another embodiment of the invention, a computer program module calculates an index of the functional relatedness of genes within each cluster, by calculating the average fraction of times that pairs of genes in the cluster are associated with the same literature unique identifier. Another computer program module randomizes the assignment of genes to clusters, then calculates the percentage of times that the index of functional similarity could have occurred by chance.

15. Please replace the text beginning on page 12, line 13:

Output data are accumulated and presented concerning the above-mentioned clusters, words and phrases, as well as the indices of functional relatedness of genes within clusters.

with the following amended text:

Output data are accumulated and presented concerning the above-mentioned clusters, words and phrases.

16. Please delete the following text beginning on page 12, line 18:

(1) The prior art clusters only data that are obtained directly from the hybridization of cDNA to microarrays, without prior transformation other than that intended to correct for (or censor) noise and other errors introduced by the measurement process, e.g., the subtraction of unwanted background that is present in images of microarrays, or the normalization of different microarray images so that reference spots have the same value in all images. An object of the present invention is to mathematically transform the error-corrected microarray data in such a way that subsequent clustering will organize genes

represented on the microarray into groups, based on mathematical models of the mechanism of the co-regulation of genes in clusters. In particular, the present invention extracts from the error-corrected microarray data estimates of the mRNA synthesis and degradation rates for each gene, which may be clustered (by any general-purpose clustering method) for purposes of explicitly identifying genes that have similar mRNA synthesis rates, which would reflect their induction by the same transcription factors. (2) The prior art provides no methods for obtaining quantitative indices for judging the quality of microarray clustering results, independent of the microarray data themselves, other than those involving databases that have already clustered genes into predetermined functional classes, e.g., the Martinsreid Institute of Sciences functional classification scheme database for yeast.

17. Please replace the text beginning on page 13, line 13:

An objective of the present invention is to generate quantitative indices about

with the following amended text:

An objective of the present invention is to generate words or phrases that describe

18. Please replace the text beginning on page 13, line 17:

An advantage of such quantitative indices

with the following amended text:

An advantage of such descriptors

19. Please replace the text beginning on page 13, line 23:

A yet further advantage of the invention is that it provides an automatic method for identifying the relevant literature for purposes of automatic analysis of the quality of clustering.

with the following amended text:

A yet further advantage of the invention is that it provides an automatic method for identifying the relevant literature, for purposes of automatic generation of key words or phrases for each cluster.

20. Please replace the text beginning on page 14, line 2:

A still further advantage of the invention is that for genes associated with a cluster, it provides a ranking of the importance of those genes on the basis of the relevance of text in literature about the set of genes in a cluster. A further advantage of the invention is that it ranks the relatedness of a cluster to all the other clusters, on the basis of the similarity of text in literature about genes in the clusters.

with the following amended text:

A still further advantage of the invention is that for each cluster, the invention ranks key words or phrases according to their importance in characterizing the cluster and distinguishing the cluster from other clusters, in which the ranking is performed after the calculation of a numerical weighting factor.

21. Please delete the following text beginning on page 14, line 11:

Fig. 2 is a block diagram of an alternate embodiment for the system and computer program product for analyzing microarray data.

Fig. 3 is a graph showing output from the system and computer program in Fig. 1, providing an example in which the system finds good evidence that a cluster's members are functionally related to one another (Cluster D in IYER et al., supra).

Fig. 4 is a graph showing output from the system and computer program in Fig. 1, providing an example in which the system finds no evidence that a cluster's members are functionally related to one another (Cluster B in IYER et al., supra).

Fig. 5 is a graph showing output from the system and computer program of Fig. 1, providing an example in which clustering was performed on microarray data described in IYER et al., supra, using the estimated mRNA synthesis rate method, in which the system finds evidence that a cluster's members are functionally related to one another.

Fig. 6 is a graph showing output from the system and computer program of Fig. 1, with data and analysis the same as in Fig. 5, except that artificial noise was added by the system to the microarray data before analysis, where the noise had a coefficient of variation of 35%.

Fig. 7 is a graph of simulated microarray time series data involving 100 microarray spots (genes), consisting of 10 sets (clusters) of 10 genes, with each gene in a set having the same mRNA synthesis function but a different mRNA degradation function.

Fig. 8 is a graph showing the results of applying the system's method for estimating mRNA synthesis rates, applied to the data shown in Fig. 7, revealing the 10 clusters.

22. Please replace the text beginning on page 17, line 3:

Installation of DJGPP is necessary in order to install the text modeling program module **(126)**, keyword identification module **(128)**, and text classification module **(130)**.

with the following amended text:

Installation of DJGPP is necessary in order to install the text modeling program module **(126)**, keyword identification module **(128)**, and associated text classification software.

23. Please replace the misspelled word "inititiation" on page 18, line 2, with the amended word "initiation".

24. Please replace the text beginning on page 24, line 9:

After downloading all of the Omim Web page files corresponding to the Omim numbers that had been associated with spots on the microarray, the UID Identification module **(118)**

with the following amended text:

After downloading all of the Omim Web page files corresponding to the Omim numbers that had been associated with spots on the microarray, and storing them in the Omim Web Pages section **(146)** of the Data Repository **(138)**, the UID Identification module **(120)**

25. Please replace the text beginning on page 24, line 15:

Then, the UID Identification module **(118)**

with the following amended text:

Then, the UID Identification module **(120)**

26. Please delete all of the text between the sentence on page 29, line 20, beginning with:

> A special option is also available for the clustering of microarray data that are in the form of a time series,

and ending the deletion of text with the sentence on page 36, line 7:

> At that point, the synthesis function for each gene in a cluster is taken to be its centroid, with an uncertainty determined by the range of functions in the cluster.

27. Please move and insert text to immediately follow the following sentence on page 38, line 7:

> The information that is provided automatically by the Keyword Identification Module **(128)** is a list of words for each cluster, sorted in descending order according to the numerical weights calculated by a classification algorithm.

by moving and inserting the following text that begins on page 41, line 23:

> The default method that the program rainbow uses for classification is the Naive Bayes method, otherwise known as Evidence Classification or simply the Bayes method. This method, along with applications related to text classification, is explained in Chapter 6 of MITCHELL, Machine Learning, McGraw-Hill (1997). The User at the User Interface **(106)** may also specify as an option that a different method be used by the computer program 'rainbow' to perform the classification, including support vector machines (svm), term frequency- inverse document frequency (tfidf), probabilistic indexing (prind), maximum entropy (maxent), k-nearest neighbors (knn), EM algorithm (em), Dirichilet kernel (dirk), and Active Learning (active). These options are then requested by the Text Modeling Module **(126)** by issuing a
>
> system("...--method=METHOD...")
>
> command for the indicated options to be performed by the rainbow computer program, where METHOD is one of the methods indicated above in parentheses.

28. Please begin the following text on page 38, line 10, as a new paragraph:

The following system function, written in the C programming language, is used by the Keyword Identification Module **(128)** to generate the word lists:

29. Please replace the text on page 39, line 2:

Examples of such word lists are given in Tables 1, 2, 3, and 4.

with the following amended text:

Examples of such word lists are given in Tables 1 and 2.

30. Please delete all of the text between the sentence on page 39, line16, beginning with:

Upon completion of these steps by the Keyword Identification Module **(128)**

and ending the deletion of text with the sentence on page 41, line 21:

The confusion matrix data are also stored in the Text Classification Data section **(160)** in the Data Repository **(138)**.

31. Please delete all of the text between the sentence on page 42, line 13, beginning with:

Upon completion of these steps by the Text Classification Module **(130)**, the Process Control Module **(116)** initiates operation of the Cluster Randomization Module **(132)**.

and ending the deletion of text with the sentence on page 44, line 16:

Thus, if text about cluster i is frequently misclassified as text about cluster j, this indicates that the genes in clusters i and j have functions in common.

32. Please replace the text beginning on page 44, line 19:

Upon completion of these steps by the Data Summarization Module **(134)**, the Process Control Module **(116)** initiates operation of the Data Output Module **(136)**. It displays the key words or phrases,

with the following amended text:

Upon completion of these steps, the Process Control Module **(116)** initiates operation of the Data Output Module **(136)**, which displays the key words or phrases that were generated in accordance with the rainbow options that had been selected.

33. Please delete the following text beginning on page 44, line 21:

> and it also provides the results of the figure of merit calculations and the Kolmogorov-Smirnov statistical tests for each of the clusters. Finally, it provides a graphical display of the text classification results, which gives an indication of the extent to which, on average, the text about each gene within a cluster resembles that of other genes within a cluster, as compared with that of genes selected at random from all the clusters. Examples of the format of that graphical display are given in Fig. 3 and Fig. 4.

34. Please delete the section of text entitled **DESCRIPTION OF ALTERNATE EMBODIMENTS** beginning on page 45, line 5, and ending with the following sentence on page 49, line 10:

> Those percent coupling scores are then stored in the Summarized Data section **(162)** of the Data Repository **(138)** for subsequent display by the Data Output Module **(136)**.

35. Please replace the text "IYER et al.,," on page 49, line 14, with the text "IYER et al.," so as to remove the extraneous comma.

36. Please delete the section of text beginning with the following on page 51, line 6:

> In a second analysis, the Clustering Module **(124)** in Fig. 1 also clustered these 517 accession numbers,

and end deletion of the section of text with the following sentence on page 51, line 23:

> The coefficient of variation for the the added noise was 30%.

37. Please delete the following text beginning on page 52, line 6:

> For the data clustered by the Clustering Module **(124)** itself after estimating mRNA synthesis rates, examples of the list of Keywords characterizing the clusters are shown in Tables 3 and 4. Table 3 gives the keywords for a cluster when no noise was added to the microarray data before processing it. Table 4 gives the keywords for the comparable cluster, produced when noise was added to the microarray data before processing.

38. Please delete Tables 3 and 4, beginning on page 55, line 1.

39. Please delete the section of text beginning with the following sentence on page 57, line 13:

> The Text Classification, Cluster Randomization, and Data Summarization Program Modules (**130**, **132**, and **134**, respectively) then processed the data as described in the Preferred Embodiment.

and end deletion of the section of text with the following sentence on page 65, line 3:

> These results are in agreement with the results shown in Figs. 3 and 4, which also indicated that the accession numbers (genes) associated with Cluster D are functionally related to one another, whereas the accession numbers (genes) associated with cluster B are not.